2 LITERATURE REVIEW

This chapter presents a survey of theories, algorithms, techniques and methods used for estimating O-D matrices. The literature review begins with a discussion of general historical developments of the subject. A variety of models for establishing O-D trip tables based on traffic volume are then addressed. The Equilibrium-based model and Linear Programming (LP) model, on which this research is based, are then studied in some detail. A short summary is provided in the last section that evaluates the different models. Typically, transportation models can be divided into dynamic and static steady-state representations. The results on static O-D trip tables are only considered in this survey, since this is the emphasis of the present research effort.

2.1 Conventional Analysis and Quick Response Models

Before the 1970s, O-D trip tables were obtained via statistical surveys, such as home interviews, license plate surveys and roadside surveys. The methods that use such survey data to determine real trip distributions are now called *conventional analysis*. The first large-scale cordon count (O-D table) was conducted in Chicago in 1916 (Easa 1993a). Prior to World War II, information on the distribution of urban traffic was obtained using roadside interviews. Since surveys were conducted through sampling, it is impossible to determine the real trip information. Furthermore, with the evolution of society and with rapid changes in transportation demands, these surveys became harder to perform, and expensive with respective to time, manpower, money and effort. Another drawback of conventional analysis is that, as the land-use changes, these data soon become out-dated. In addition, in most cases, an assignment of this matrix to the network cannot reproduce observed flows. Three main types of models are considered in the conventional analysis: Fratar models, opportunity models and gravity models. These are evaluated by Easa (1993a), and the reader is referred to this paper for a detailed illustration of characteristics, advantages and limitations of these models.

The emphasis on transportation system management in the early 1970s increased the need for studying small urban areas in detail. Some cheaper and quicker-response theory and methods for synthesizing trip tables from more conveniently available information have been developed since then. According to the purpose of these models, they can be classified into the following subgroups: National Cooperative Highway Research Program (NCHRP) Simplified Techniques, Traffic Count-Based Models, Self-Calibrating Gravity Models, Partial Matrix Techniques, models using GIS data, Heuristic Methods and Facility Forecasting Techniques. There are also some special application models, for example, freeway trip distribution, pedestrian trip distribution and special purpose trip-distribution models. Examples of special purpose trip-distribution models include choice models (employing individual travelers instead of the zones as the unit of observation), continuous models (that ignore the zones altogether when the changes in land-use patterns are small), simultaneous models (that simultaneously analyze trip distributions and other planning steps). We do not discuss the differences among these subgroups in detail. The interested reader is recommended to refer to the review papers by Easa (1993b), Sivanandan (1991), and Sivanandan *et al.* (1996) and the references listed therein. For a review and evaluation of some microcomputer programs in quick-response analysis, the reader is referred to Easa's paper.

2.2 Traffic Count-Based Models

In reality, the problem on a given network is a dynamic (time-dependent) system. The analytical traffic flow process on such networks can be simplified to study some density function that satisfies certain differential or difference equations of the *continuation type* (Lyrintzis *et al.* 1994, Zhang *et al.* 1995), or can be treated by dynamic programming (DP) approaches (Janson 1995). In practice, however, this dynamic system is very hard to study to obtain a desired solution. In order to derive O-D trip tables from count-based information, the traffic flow is usually considered in its static (time-independent) or stationary state. There is some evidence to support such an idea. First, the study of O-D trip tables is motivated by the purpose of reducing the congestion problem. In the case of congestion, however, the distribution of traffic flow is dominated by the *user equilibrium principle* discussed in the next section. Moreover, it can be shown that the flow pattern is independent of time if the system satisfies an equilibrium condition. Second, if the system is considered during a very short time period, it can be approximated as a static system. In addition, the stationary treatment of dynamic systems can be thought as a first step toward problem simplification, and frequently, this reveals a more detailed structure of the problem.

Among all types of easily derived data, traffic counts (link volumes in a network) perhaps contains the most important information about O-D distributions. A variety of analytic models have been developed to establish O-D trip tables based on traffic counts along with other information. Based on the theory (or principle, or hypothesis) used, these models may be divided into the following types (Easa 1993b, CTR 1995).

I. Gravity-Based Models

These models are sometimes called Parameter Calibration models, and represent the original idea of establishing trip distributions. In these models, the entries of the O-D matrix are assumed to be a function of the traffic count and other parameters. Regression techniques and the flow conservation law are applied to calibrate the parameters such that the differences between observed volumes and established volumes are minimized. The models are divided into linear (Low 1972, Holm *et al.* 1976, Gaudry and Lamarre 1978, Smith and McFarlane 1978) and nonlinear (Rolillhard 1975, Hogbag 1976) regression models.

II. Equilibrium Models

These models are based on the principle of user optimization of traffic flow, called the "Equilibrium Principle" or "Wardrop's Principle" (Wardrop 1952). These include LINKOD (Nguyen 1977a-b, Gur 1980), SMALD (Kurth *et al.* 1979), and LP (Sivanandan 1991, Sherali *et al.* 1994a-b). The equilibrium-based models will be discussed in detail in the following sections.

III. Entropy Models

Minimum Information and Maximum Entropy models are included in this group and can be converted to a type of gravity models. In these models, the probability of a particular trip distribution occurring is assumed to be proportional to the number of the states (entropy or disorder) of the system. The derived O-D table is purported to be the most likely one that is consistent with information such as length and free speed of the links contained in the link flows. The pioneers of these models, ME2, are Wilumsen (1978) and Van Zuylen (1978, 1979). Many improvements and combinations with other theory have been conducted since then, and a great deal of testing on these models has been performed. These results have been summarized in the review papers of Easa (1993b) and CTR (1995).

IV. Statistical Models

These models take into account inaccuracies on the observed O-D flows, row and column sums and traffic counts. This group includes the constrained generalized least squares (CGLS) model (McNeil 1983), and constrained maximum-likelihood (CML) models (Geva 1983, Spiess 1987, Walting and Maher 1988, Walting and Grey 1991). Another model, called MEUSE, standing for Matrix Estimation Using Structure Explicitly (Bierlaire and Toint 1995), which uses both historic data and parking data as input, can be partially included in this subgroup. Statistical models are not as popular in practice as compared with equilibrium and entropy models.

V. Neural Network Models

Muller and Reinbardt (1990) introduced the neural network approach to determine O-D trip table from traffic counts. In a loose sense, this approach is based on the concepts derived from research into the nature of the brain. The procedure in this approach includes “learning” and “optimization” components. The model may be mathematically described as a directed graph with three characteristics. 1) A state of variable associated with each node; 2) a weight assigned to each link; and 3) a transfer function defined for determining the state of each node as a function of its bias and weights of its incoming links. Yang, Akiyamma and Sasaki (1992) adopted a feed-forward neural network for synthesizing O-D flow for a four-way intersection and a short freeway segment. Chin, Hwang and Pei (1994) described a neural network model for generating O-D information from flow volumes. There still exists a need for further testing these models.

VI. Fuzzy Weight Models

Instead of the “all or nothing” assumption made in most models, Fuzzy Weight approaches apply some kind of “fuzziness” to the link data (Xu and Chan 1992a-b). Fuzziness indicates probability, but is quite different in nature. Different types of fuzziness have been tested on the network of the Eastern Highway Corridor. The model is relatively new, and additional case studies and experimentation are recommended for evaluation.

2.3 Equilibrium-Based Models

The User Equilibrium Principle, also known as Wardrop's (first) principle or user-optimal assignment principle, was originally used to guide the traffic flow assignment process. It requires that all routes having positive flows between any O-D pair should have equal traffic cost, and this cost must not exceed the traffic cost on any other unused route between this O-D pair. Mathematically, estimating the O-D trip tables from link volumes can be thought of as an inverse problem of transportation assignment. Therefore, the equilibrium-based O-D matrix estimation problem may be described as determining an O-D matrix such that, when this matrix is user-optimally assigned to the network, it reproduces the observed O-D travel times. Since the correspondence between equilibrium link flow patterns and equilibrium $\mathrm{O-D}$ travel times for the standard equilibrium problem is "one-to-one" (Yang *et al.* 1994), it consequently reproduces the observed link flows.

Nguyen (1977) exploits Wardrop's user-equilibrium principle for route choice and formulates a deterministic network equilibrium approach. He shows that, if the set of observed link flows is at equilibrium, the solution matrix can be found by solving the following optimization problem with respect to variables v and t

Nguyen's Model 0 (NM0)

$$\operatorname*{Min}_{v,t} F(v,t) = \sum_{a\in A}\int_0^{v_a} f_a(x)\,dx + \sum_i \sum_j \bar{u}_{ij} t_{ij} \qquad [2.1a]$$

subject to:

$$\sum_r h_{r_{ij}} = t_{ij}, \qquad \forall\, i, j \qquad [2.1b]$$

$$\sum_{r_{ij}} d_{a,r_{ij}}\, h_{r_{ij}} = v_a, \qquad \forall\, a \qquad [2.1c]$$

$$\sum_{i\,j} \bar{u}_{ij}\, t_{ij} = \sum_{a} f_a(\bar{v}_a)\bar{v}_a \qquad [2.1d]$$

$$h_{r_{ij}} \geq 0, \qquad t_{ij} \geq 0.$$

Here $f_a(x)$ is the volume-delay function for link a, t_{ij} is the travel demand between origin i and destination j, v_a is the assigned flow on link a, $h_{r_{ij}}$ is the flow on route r connecting origin i and destination j,

$$d_{a,r_{ij}} = \begin{cases} 1 & \text{if link } a \text{ is on route } r \text{ between } i \text{ and } j \\ 0 & \text{otherwise,} \end{cases}$$

$\bar{v}_a$ is the observed link flow, and $\bar{u}_{ij}$, the observed inter-zonal accessibility (travel time on any used route between zone i and zone j due to the equilibrium assumption).

As shown by Nguyen (1984), for each link $a \in A$, the problem *NM0* has a unique optimal solution v_a^*, because the objective function is strict convex in v. The optimal solution v_a^* is consistent with the observed link flow $\bar{v}_a$, $a \in A$, if the observed link flow pattern is at equilibrium. However, the problem *NM0* may not have a unique O-D matrix solution because that it is not strictly convex in the variable t (Sheffi 1985). Namely, there are many possible O-D matrices that produce the same set of observed O-D travel times of observed link flows when assigned to the network.

Nguyen's theoretical model was operationalized during the course of a Federal Highway Administration Project in which the LINKOD system of models were developed (Turnquist and Gur 1979; Gur *et al.* 1980). The LINKOD system is comprised of two major components: SMALD and ODLINK. SMALD (Kurth *et at.* 1979), is a small area trip distribution model that determines a trip table for a sub-area. This table is used to overcome the under-specification problem of Nguyen's formulation and has been extensively tested and verified by Han, Dowling, Sullivan and May (1981), Han and Sullivan (1983), and Dowling and May (1984). To obtain a unique O-D matrix, a target O-D matrix $\bar{t}_{ij}$ is assumed to be available. Different criteria have been suggested for choosing among all the O-D matrices that produce the observed link flows. In the LINKOD model (Gur 1980), the estimation problem is formulated as

NM1

$$\min \sum_{ij} (t_{ij} - \bar{t}_{ij})^2 \qquad [2.2]$$

subject to t being optimal to *NM0*.

Another model is suggested by Jornsten and Nguyen (1980) that chooses the most likely matrix among the optimal solution set by solving the following optimization problem,

NM2

$$\max \sum_{ij} - t_{ij} \left\{ \log\left(\frac{t_{ij}}{\bar{t}_{ij}} \right) - 1 \right\} \qquad [2.3]$$

subject to t being optimal to *NM0*.

In the existing literature, two requirements, namely, to reproduce the observed O-D travel times and to reproduce the observed link flows, have been used. These two requirements are equivalent when the observed network is in equilibrium. The equivalence relationship, however, will not hold when the set of observed link flows is not an equilibrium solution (Yang *et al.* 1994).

Both NM1 and NM2 have a bi-level programming structure that poses computational difficulties for large-scale networks (Fisk 1989, Fisk & Boyce 1983). LeBlanc and Farhangian (1982) suggested a partial dualization method for solving NM1 (Sheffi 1985). The partial dualization method involves updating a Lagrange multiplier by iteratively solving a Lagrangian minimization problem. A similar method for solving

NM2 was also suggested by Jornsten and Nguyen (1980) and Nguyen (1984). Both of these methods require iteratively solving program NM0, and are computational demanding. Yang *et al.* (1992b) have shown how to integrate existing methods such as the generalized least squares technique with an equilibrium traffic assignment approach using a Stackelberg leader-follower optimization model. An attempt is made for uncertainties in both the target O-D matrix and in the traffic link counts, and a heuristic solution method is proposed because of the inherent difficulty in solving moderate to large sized problems of this type.

It has been shown by Yang *et al.* (1994) that Nguyen's bi-level optimization models can be transformed into a single convex program. When the observed link flow pattern is in equilibrium, the original model is demonstrated to be equivalent to a reduced system of linear equations. By exploiting the properties of the system's feasible region, simpler methods, such as a least squares technique, can be used to obtain an O-D matrix that, when user-optimally assigned to the network, will reproduce the observed link flows.

The models mentioned above are based on improving and modifying Nguyen's systems. Other models based directly on the user-optimal principle are also developed. One of them is to apply Linear Programming (LP) theory to estimate $\mathrm{O-D}$ trip tables from the observed link flows (Sherali *et al.* 1994a, Sivanandan 1991). The detailed studies on LP models are presented in the next section.

2.4 Linear Programming Models

Linear programming theory and technique have been successfully applied to various transportation problems almost since its early beginning. A famous example is given by Dantzig [1951] to adapt his simplex method to solve (Hitchcock's) transportation problem. The terminology, such as transportation/assignment problems, and allocation problems, have become a standard in these contexts since then (Bazaraa *et al.* 1990).

Linear programming methods were first used to study O-D distributions in the 1970s (Colston 1970). The developed approaches, unfortunately, were not successful in practice. An alternative linear programming approach has been proposed recently (Sivanandan 1991, Sherali *et al.* 1994a-b) based on a non-proportional assignment and user-equilibrium principle. The basic idea and technique used in this modeling approach may be described as follows.

Consider a road network. Let N be the set of nodes including centroids, A be the set of corresponding directed links or arcs, and let OD be the set of O-D pairs that comprise the trip table to be estimated. Let x_{ij} be the estimation of flow originating at i and destined for j, and $\bar{f}_a$ be the observed link flow for each link $a \in A$. Let p_{ij}^k, $k = 1, \cdots, n_{ij}$, represent all the n_{ij} paths between each O-D pair $(i, j) \in OD$, and x_{ij}^k be the contribution of x_{ij} to the path p_{ij}^k for each $k = 1, \cdots, n_{ij}$. Assume that the link travel time/cost $c_a(v_a)$ is a (strictly) increasing function of flow v_a, let $c_{ij}^k = \bar{c} \cdot p_{ij}^k$ denote the cost on route k between O-D pair (i, j) for $k = 1, \cdots, n_{ij}$, and let $c_{ij}^* = \min\{c_{ij}^k, k = 1, \cdots, n_{ij}\}$. Then based on the user-optimal principle, the following linear programming model may be constructed.

Minimize
$$\sum_{(i,j)\in OD} \sum_{k=1}^{n_{ij}} \hat{c}_{ij}^k x_{ij}^k \qquad [2.4a]$$

subject to
$$\sum_{(i,j)\in OD} \sum_{k=1}^{n_{ij}} (p_{ij}^k)\, x_{ij}^k = \bar{f} \qquad [2.4b]$$

$$x \geq 0 \qquad [2.4c]$$

where

$$\hat{c}_{ij}^k = \begin{cases} c_{ij}^k & \text{if } k \in K_{ij} \\ M_1 c_{ij}^k & \text{if } k \notin K_{ij} \end{cases} \qquad [2.4d]$$

and where $M_1 > 1$ is a constant, and K_{ij} is the observed inter-zonal accessibility defined by $K_{ij} = \{k \in \{1, \cdots, n_{ij}\} : c_{ij}^k = c_{ij}^*\}$, for each $(i, j) \in OD$.

It has been shown by Sherali et al. that the observed link flow pattern is at an equilibrium if and only if the objective value of [2.4] is equal to $\overline{C}_{total} = \sum \overline{c}_a \bar{f}_a$, the total observed system cost. In order to take the internal inconsistencies of input data and the prior trip table into account, some penalized terms are added to the objective function and the following LP and LP(TT) models are considered, respectively.

LP:

Minimize $$\sum_{(i,j)\in OD} \sum_{k=1}^{n_{ij}} \hat{c}_{ij}^k x_{ij}^k + Me \cdot (y^+ + y^-) \qquad [2.5a]$$

subject to $$\sum_{(i,j)\in OD} \sum_{k=1}^{n_{ij}} (p_{ij}^k)\, x_{ij}^k + (y^+ - y^-) = \bar{f} \qquad [2.5b]$$

$$x \geq 0,\ y^+ \geq 0,\ \ y^- \geq 0. \qquad [2.5c]$$

LP(*TT*):

Minimize $$\sum_{(i,j)\in OD} \sum_{k=1}^{n_{ij}} \hat{c}_{ij}^k x_{ij}^k + Me \cdot (y^+ + y^-) + M_\sigma \sum_{(i,j)\in \overline{OD}} (Y_{ij}^+ + Y_{ij}^-) \qquad [2.6a]$$

subject to $$\sum_{(i,j)\in OD} \sum_{k=1}^{n_{ij}} (p_{ij}^k)\, x_{ij}^k + (y^+ - y^-) = \bar{f} \qquad [2.6b]$$

$$\sum_{k=1}^{n_{ij}} x_{ij}^k + (Y_{ij}^+ - Y_{ij}^-) = Q_{ij} \qquad \forall (i, j) \in \overline{OD} \qquad [2.6c]$$

$$x \geq 0,\ y^+ \geq 0,\ \ y^- \geq 0,\ Y^+ \geq 0,\ Y^- \geq 0. \qquad [2.6d]$$

Here, $\overline{OD} \subseteq OD$ is a subset of O-D pairs for which a partial prior (target) trip table information is available, and M and M_σ are some positive penalty parameters. The authors investigated on how large these penalty coefficients should be for the purpose of achieving desired solutions. Instead of using the (standard) Simplex Method, a Column Generation Algorithm (CGA) is developed for solving LP(TT), in order to reduce the computational effort for practical problems. The models have been tested on a real network of Northern Virginia.

In practice, it is commonly the case that not all of the link volumes of the network are available. The LP models have been improved to take care of the case of missing volume (Sherali *et al.* 1994b). The idea used in this situation is to update the travel time/cost by solving some linear and nonlinear programming subproblems iteratively. Both the original and improved versions have been tested and evaluated on some real networks (Sivanandan *et al.* 1996). Some of these test results will be presented in Chapter 6.

2.5 Summary of Literature Review

Different approaches have been investigated for estimating O-D trip tables in the last several decades. Conventional analysis is very expensive in practice. It is hard to reproduce the observed flows using these techniques, and the trip tables often become outdated. As a consequence, quick response models become of great relevance in transportation planning.

Among quick response models, the one based on traffic counts is very popular and pervasive. The problem is beset with a great deal of complexities, and various approaches have been employed to overcome them.

Gravity-based models require considerable data, and are relatively more likely to have their results become outdated. This makes these models unattractive. LINKOD type models incorporate the desired equilibrium assignment concept, but their nonlinear nature leads to the issue of excessive computational effort for deriving acceptable solutions to practical problems. The entropy-based models pose restrictions on data, give little weight to prior information, and need refinements for incorporating the equilibrium principle. Statistical models take into account the stochastic nature of the data and the problem. However, they have not been adequately tested. In addition, the stochastic theory used itself sometimes makes the problem more complicated for practical purposes. Both neural network and fuzzy set approaches still need to be verified regarding their practical viability. Moreover, in-depth theoretical studies are themselves needed before these approaches can be justified for use in practice.

Linear programming models and algorithms have been widely used in various applications, including transportation and assignment problems. Using this approach to estimate the O-D trip matrix from link volumes, however, is relatively new. The approaches have some advantages such as a simpler formulation, and for the case of all link volumes being available, an established theory guaranteeing finite convergence. On the other hand, the approach approximates the random nature of the data and the problem, and the resulting O-D table often has many zeros because of the "extreme point" optimality principle. (This is somewhat alleviated when using prior trip table information.) Moreover, in the case of missing volumes, the approach iteratively updates the travel cost on missing data links and then minimizes the total cost, which leads to a problem of excessive computational effort.

Therefore, there is a need to improve the LP model so that it can take care of inaccuracies in input data in practice, interpret both user behavior and *user-optimal* principles in a reasonable manner, and reduce the computational effort in generating practical acceptable solutions. With this motivation, a linear programming model and approach is proposed in the next chapter.